

University of New Hampshire

University of New Hampshire Scholars' Repository

Molecular, Cellular and Biomedical Sciences
Scholarship

Molecular, Cellular and Biomedical Sciences

12-6-2007

Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography

Philippe Normand
Université de Lyon

Pascal Lapierre
University of Connecticut, Storrs

Louis S. Tisa
University of New Hampshire, Durham, louis.tisa@unh.edu

Johann P. Gogarten
University of Connecticut, Storrs

Nicole Alloisio
Université de Lyon
For more and additional works at: https://scholars.unh.edu/mcbs_facpub

Comments

This is an article published by Cold Spring Harbor Laboratory Press in *Genome Research* in 2007, available online:
See next page for additional authors
<https://dx.doi.org/10.1101/gr.5798407>

Recommended Citation

Normand, P., P. Lapierre, L. S. Tisa, J.P. Gogarten, N. Alloisio, B. Cournoyer, C. Lavire, J. Marechal, P. Pujic, J. P. Gogarten, Y. Huang, J. Mastronunzio, D. Bickhard, C. Bassi, T. Rawnsley, J. Niemann, M. P. Francino, A. Lapidus, M. Martinez, E. Goltsman, G. Perriere, C. Medigue, N. Choisne, A. Couloux, S. Cruveiller, L. Labarre, Z. Rouy, D. Vallenet, C. T. Cong Y, N. Demange, B. Mullin, O. Kopp, Y. Wang, J. Tomkins, A. Berry, A. Sellstedt, F. Tavares, C. Valverde, L. Wall, and D. R. Benson. 2007. Genome characteristics of three facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography. *Genome Research* 17:7-15 (<http://www.genome.org/doi/10.1101/gr.5798407>).

This Article is brought to you for free and open access by the Molecular, Cellular and Biomedical Sciences at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Molecular, Cellular and Biomedical Sciences Scholarship by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact Scholarly.Communication@unh.edu.

Authors

Philippe Normand, Pascal Lapierre, Louis S. Tisa, Johann P. Gogarten, Nicole Alloisio, Emilie Bagnarol, Carla A. Bassi, Alison M. Berry, Derek M. Bickhart, Nathalie Choisne, Arnaud Couloux, Benoit Cournoyer, Stephane Cruveiller, Vincent Daubin, Nadia Demange, Maria Pilar Francino, Eugene Goltsman, Ying Huang, Olga R. Kopp, and Laurent Labarre

Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography

Philippe Normand,¹ Pascal Lapierre,² Louis S. Tisa,³ Johann Peter Gogarten,² Nicole Alloisio,¹ Emilie Bagnarol,¹ Carla A. Bassi,² Alison M. Berry,⁴ Derek M. Bickhart,² Nathalie Choisne,^{5,6} Arnaud Couloux,⁶ Benoit Cournoyer,¹ Stephane Cruveiller,⁷ Vincent Daubin,⁸ Nadia Demange,⁶ Maria Pilar Francino,⁹ Eugene Goltsman,⁹ Ying Huang,² Olga R. Kopp,¹⁰ Laurent Labarre,⁷ Alla Lapidus,⁹ Celine Lavire,¹ Joelle Marechal,¹ Michele Martinez,⁹ Juliana E. Mastronunzio,² Beth C. Mullin,¹⁰ James Niemann,³ Pierre Pujic,¹ Tania Rawnsley,³ Zoe Rouy,⁷ Chantal Schenowitz,⁶ Anita Sellstedt,¹¹ Fernando Tavares,¹² Jeffrey P. Tomkins,¹³ David Vallenet,⁷ Claudio Valverde,¹⁴ Luis G. Wall,¹⁴ Ying Wang,¹⁰ Claudine Medigue,⁷ and David R. Benson^{2,15}

¹Université de Lyon, Unité Mixte de Recherche, Centre National de la Recherche Scientifique (UMR CNRS), 5557 Ecologie Microbienne, IFR41 Bio Environnement et Santé, Université Lyon I, Villeurbanne 69622 cedex, France; ²Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut 06279, USA; ³Department of Microbiology, University of New Hampshire, Durham, New Hampshire, 03824, USA; ⁴Department of Plant Sciences, University of California, Davis, California 95616, USA; ⁵INRA, Institut National de la Recherche Agronomique–Unité de Recherche en Génomique Végétale (INRA-URGV), 91057 Evry cedex, France; ⁶Genoscope, Centre National de Séquençage, 91057 Evry cedex, France; ⁷Genoscope, CNRS-UMR 8030, Atelier de Génomique Comparative, 91006 Evry cedex, France; ⁸Bioinformatics and Evolutionary Genomics Laboratory, UMR CNRS 5558, Université Lyon I, Villeurbanne 69622 cedex, France; ⁹DOE Joint Genome Institute, Walnut Creek, California 94598, USA; ¹⁰Department of Biochemistry & Cellular & Molecular Biology and The Genome Science & Technology Program, The University of Tennessee, Knoxville, Tennessee 37996, USA; ¹¹Department of Plant Physiology, Umeå University, S-90187 Umeå, Sweden; ¹²Instituto de Biologia Molecular e Celular, Universidade do Porto, 4150-180 Porto, Portugal; ¹³Clemson University Genomics Institute, Clemson, South Carolina 29634, USA; ¹⁴Programa Interacciones Biológicas, Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal B1876BXD, Argentina

Soil bacteria that also form mutualistic symbioses in plants encounter two major levels of selection. One occurs during adaptation to and survival in soil, and the other occurs in concert with host plant speciation and adaptation. Actinobacteria from the genus *Frankia* are facultative symbionts that form N₂-fixing root nodules on diverse and globally distributed angiosperms in the “actinorhizal” symbioses. Three closely related clades of *Frankia* sp. strains are recognized; members of each clade infect a subset of plants from among eight angiosperm families. We sequenced the genomes from three strains; their sizes varied from 5.43 Mbp for a narrow host range strain (*Frankia* sp. strain HFPCcl3) to 7.50 Mbp for a medium host range strain (*Frankia alni* strain ACN14a) to 9.04 Mbp for a broad host range strain (*Frankia* sp. strain EANlpec.) This size divergence is the largest yet reported for such closely related soil bacteria (97.8%–98.9% identity of 16S rRNA genes). The extent of gene deletion, duplication, and acquisition is in concert with the biogeographic history of the symbioses and host plant speciation. Host plant isolation favored genome contraction, whereas host plant diversification favored genome expansion. The results support the idea that major genome expansions as well as reductions can occur in facultative symbiotic soil bacteria as they respond to new environments in the context of their symbioses.

[The genome sequences for *Frankia* strains Ccl3, ACN14a, and EANlpec have been submitted to GenBank under accession nos. CP000249, CT573213, and AA1100000000, respectively.]

¹⁵Corresponding author.

E-mail david.benson@uconn.edu; fax 860-486-4331.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.5798407>.

Two very different groups of bacteria can form nitrogen-fixing root nodules on angiosperms: Gram-negative proteobacteria from several families, and high Mol% G+C Gram-positive actinobacteria in the family *Frankiaceae*. Nodulating proteobacteria

have symbiotic genes (*nod* genes) subject to horizontal transfer among α - and some β -Proteobacteria (Chen et al. 1991; Young and Haukka 1996; Moulin et al. 2001). In contrast, all *Frankia* sp. strains are closely related with no evidence of dissemination of nodulating ability to related actinobacteria (Fig. 1; Normand et al. 1996; Clawson et al. 2004).

In plants, the capacity to form N_2 -fixing root nodules occupied by bacteria is retained in a single lineage of angiosperms known as the “ N_2 -fixing clade” (Soltis et al. 1995). Ten families within the Eurosid I clade have members that are nodulated (Soltis et al. 1995; Swensen 1996; Clawson et al. 2004). Only two of the families have members that associate with nodulating proteobacteria, while eight associate with *Frankia* sp. strains to form the actinorhizal symbiosis (Table 1).

Frankia strains fall into three closely related clusters. Members of each cluster have distinct host ranges (Table 1; Fig. 1). Cluster 1 strains nodulate plants in the Fagales in the Betulaceae and Myricaceae and are often referred to as “*Alnus* strains” (Normand et al. 1996). A subclade within Cluster 1 is comprised of the narrow host range “*Casuarina* strains” that under natural conditions nodulate only *Casuarina* and *Allocasuarina* species in the Casuarinaceae (Benson et al. 2004). Conversely, Cluster 3 “*Elaeagnus* strains” are considered to have a broad host range since they nodulate plants from five families in the Fagales and Rosales (Benson et al. 2004). Finally, the “*Rosaceous* strains” form Cluster 2, which is sister to the others; representatives of this cluster have not been isolated and grown in culture. Cluster 2 strains nodulate plants from four families in the Rosales and Cucurbitales (Benson et al. 2004; Vanden Heuvel et al. 2004).

To gain insight into the evolutionary trajectory followed by these closely related, yet host-range and geographically divergent, *Frankia* sp. strains, we sequenced and compared the genomes of three isolates, including a narrow host range *Casuarina* strain, a medium host range *Alnus* strain, and a broad host range *Elaeagnus* strain. The results suggest that gene deletion and duplication have occurred to different extents in the genomes during adaptation to host plants and their environments. The concept of genome contraction echoes the changes known to occur

in obligate bacterial pathogens and symbionts (Mira et al. 2001; Ochman and Moran 2001; Moran 2003), but the observation that both contraction and expansion can occur in closely related lineages of facultatively symbiotic soil bacteria in relation to host distribution has not previously been reported.

Results and Discussion

Actinorhizal plant families emerged in the late Cretaceous (~100 million years ago [Mya]) and subsequently adapted to a wide variety of environments (Magallon et al. 1999). Currently, they are globally distributed in climate zones ranging from alpine and subarctic to tropical (Fig. 2) where they add nitrogen and organic material to nutrient-poor soils (Silvester 1976). The native geographical distributions of hosts range from limited in the case of *Casuarina* sp. to broad in the case of *Morella* sp. (Fig. 2). The distribution of bacterial symbionts is obviously more difficult to assess, but numerous studies have shown some correlation with plant distribution (for review, see Benson et al. 2004).

Frankia sp. strain HFPCc13 (Cc13) represents narrow host range *Casuarina* strains commonly detected in nodules collected from casuarinas in their native Australia (Fig. 2A) and in areas of the world where casuarina trees have been planted as windbreaks or for erosion control (Simonet et al. 1999). Similar strains have not been found in soils in the absence of a suitable host, indicating that the bacteria depend on the plant for their soil propagation (Simonet et al. 1999).

Frankia alni strain ACN14a (ACN) represents *Alnus* strains that are globally distributed in soils regardless of the presence of a suitable host plant (Benson et al. 2004). This ubiquity parallels the distribution of host plants from the Betulaceae and Myricaceae that have a combined native range spanning all continents except Australia (Table 1; Fig. 2B).

Frankia sp. strain EAN1pec (EAN) represents broad host range *Elaeagnus* strains that are also globally distributed in soils with or without host plants (Benson et al. 2004). Cognate hosts are the most diverse and have the widest distribution with representatives on all continents including Australia (Table 1; Fig. 2C).

The strains used in this study have 16S rRNA gene sequences that are 97.8% identical between ACN or Cc13 versus EAN, and 98.9% identical between ACN and Cc13 (Fig. 1). This similarity level is frequently observed among bacteria from the same species (Wayne et al. 1987; Gevers et al. 2005), and is typical of the similarity levels found within the genus *Frankia* (Fig. 1; Clawson et al. 2004).

Genome characteristics

The genomes from ACN and Cc13 have been finished, and that from EAN has been rendered in a single scaffold with some gaps corresponding to regions that have proven difficult to resolve due to sequence repeats and high GC content (Table 2). Nevertheless, unlike *Streptomyces* (Bentley et al. 2002), all three genomes are circular as demonstrated directly from their sequences (Fig. 3). None of the strains have yielded independently replicating plasmids. Unlike what is observed for obligate symbionts, the coding capacity of the genomes remains quite high at 89% for ACN, 84% for Cc13, and 86% for EAN.

The most striking difference between the three genomes is their sizes, ranging from 5.43 Mb for Cc13 (4499 protein-coding sequences [CDS]), to 7.50 Mb for ACN (6786 CDS), to 9.04 Mb for

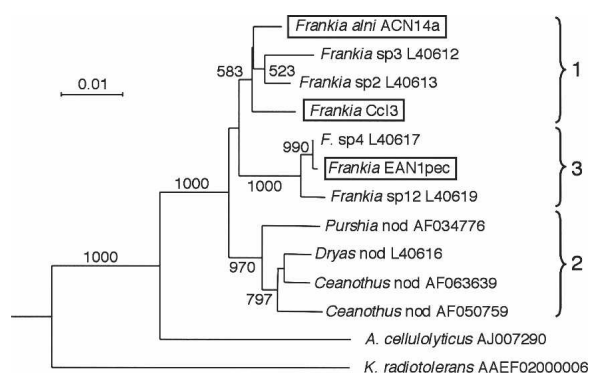


Figure 1. Neighbor-joining (Saitou and Nei 1987) phylogenetic tree calculated with ClustalX 1.83 (Thompson et al. 1997) from 16S rRNA gene sequences. Distances were corrected for multiple substitutions (Kimura 1980); otherwise, default settings were used. Numbers give bootstrap support values from 1000 bootstrapped samples. The out-group used is *Streptomyces coelicolor* (NC003888). Accession numbers for the organisms are given after the name and species number as given in Normand and Fernandez (2007). In the case of the unisolated cluster 2 frankiae, the host plant genus from which 16S rRNA gene sequences were amplified is given. Distances in the bar are in substitutions/site.

Table 1. Relationship between actinorhizal plants and groups of *Frankia* strains^a

Order ^b	Family (#nod/#gen) ^c	Genus (No. species)	Geographical distribution	<i>Frankia</i> cluster ^d	
Fagales	Betulaceae (1/6)	<i>Alnus</i> (30)	N. temperate, SA, N. Africa, Asia	1, (3)	
		<i>Comptonia</i> (1)	Eastern NA	1, 3	
	Myricaceae (3/4)	<i>Morella</i> (20)	Cosmopolitan, not Australia or Mediterranean	1, 3	
		<i>Myrica</i> (2)	Circumpolar	1 (3)	
		<i>Allocasuarina</i> (58)	Australia	1	
	Casuarinaceae (4/4)	<i>Casuarina</i> (17)	Australia	1 (3)	
		<i>Ceuthostoma</i> (1)	Malesia	ND	
		<i>Gymnostoma</i> (10)	Malaysia to W. Pacific	3	
	Rosales	Elaeagnaceae (3/3)	<i>Elaeagnus</i> (10)	Europe, Asia, NA	3
<i>Hippophae</i> (2)			Temperate Eurasia	3	
<i>Shepherdia</i> (2)			NA	3	
Rhamnaceae (6/55)		<i>Ceanothus</i> (55)	Western NA	2, (3)	
		<i>Colletia</i> (17)	Southern SA	3	
		<i>Discaria</i> (15)	Southern SA, Australia, New Zealand	3	
		<i>Kentrothamnus</i> (1)	Southern SA	3	
		<i>Retanilla</i> (4)	Southern SA	3	
		<i>Trevoa</i> (1)	Southern SA	3	
		<i>Cercocarpus</i> (6–10)	Western NA	2	
Rosaceae (4/100)		<i>Chamaebatia</i> (2)	Western NA	2	
		<i>Dryas</i> (2–3)	Circumboreal, arctic-alpine	2	
		<i>Purshia</i> (8)	Western NA	2	
Cucurbitales		Coriariaceae (1/1)	<i>Coriaria</i> (5–20)	Mexico to SA, W. Mediterranean, New Zealand, Papua New Guinea, S.E. Asia	2
		Datisceae (1/1)	<i>Datisca</i> (2)	Western NA, S. Asia	2

^aCompiled after Benson et al. (2004) and Swensen (1996).

^bAccording to the classification of the Angiosperm Phylogeny Group (APG 1998); all of these orders fall in the "Eurosoid 1" group of eudicots.

^cThe number of genera within the family is listed along with the number of genera nodulated. Not all genera within a family are capable of nodulation.

^dThe *Frankia* cluster refers to the clusters of *Frankia* strains in Figure 1 typically found in root nodules of each plant genus. 1, alder and casuarina strains; 3, elaeagnus strains; 2, rosaceous strains; ND, not determined; (3) indicates rare detection of an elaeagnus strain in a nodule or in surface layers of a nodule (Benson et al. 2004).

(NA) North America; (SA) South America.

EAN (7976 CDS) (Table 2). On a total DNA basis, this range is the largest reported for any group of free-living prokaryotes related at the 98%–99% 16S rRNA sequence level described to date. The smallest genome belongs to the narrow host range and geographically limited representative CcI3, and the largest is from strain EAN, which belongs to the broadest host range group. This size correlation raises the hypothesis that genome size and content is driven by the host range and biogeography of the symbiosis. We addressed this hypothesis through comparative analysis of the genomes' contents and structures to determine how the disparate sizes have come about.

The genome maps shown in Figure 3 indicate that the patterns of synteny are quite similar, with synteny decreasing as the terminus of replication is approached, corresponding to a high degree of gene rearrangement, duplication, or deletion in this region. Indeed, much of the size differences can be accounted for by expansion in this area of the genomes of EAN and ACN. Genes related to symbiosis shown in Figure 3 include those encoding nitrogenase (*nif*), uptake hydrogenase (*hup*), and squalene hopane cyclase (*shc*) involved in bacteriohopane biosynthesis. Only one gene similar to the common nodulation genes in rhizobia has been found in each strain, but with a general function prediction and relatively low BLAST scores (the product of FRAAL4911, annotated as a chitin deacetylase, resembles NodB from *Rhizobium* sp. N33 with a score of $3e^{-34}$ but also resembles similar proteins from many other organisms, mainly *Bacillus* sp.). Little is known about the genetic basis of the actinorhizal symbiosis, but it is clearly very different from that known to exist among the nodulating proteobacteria.

IS elements and prophage

Insertion elements, transposases, integrated phage, and plasmids tend to reflect the degree of plasticity of genomes (Frost et al. 2005). In the three *Frankia* strains, integrases plus transposases comprise ~0.6% (46) of the ACN CDS, 4.1% (187) of the CcI3 CDS, and 3.4% (269) of the EAN CDS (Table 3). Such elements tend to cluster in regions where there is loss of synteny between genomes (Fig. 3). CcI3 has a higher average density of mobile elements (34 per Mb) than EAN (30 per Mb), and both have a much higher density than ACN (6 per Mb). Indeed, of the 33 transposase genes identified in ACN, all but four are found in the same context in CcI3, and all but six are in EAN, suggesting that such genes in ACN have been inactive for at least the 55 million years since the genomes diverged (Normand et al. 1996; Clawson et al. 2004). Examining regions surrounding phage integrases, we estimate that prophage contribute 0.4%, 11.7%, and 7.1% of the CDS of ACN, CcI3, and EAN, respectively. Overall, EAN and CcI3 have had far more dynamic genomes than ACN in their recent history, and this plasticity, conferred by IS elements and phage moving into and out of the genomes, may in large part have driven the size differences observed.

Gene deletions

To examine how the three genomes have evolved to have such different sizes, we used the order of divergence of the three strains (Fig. 1) to estimate how the genome size differences reflect gene deletion, duplication, and acquisition. Using the approximation of 50 Myr per 1% divergence in the 16S rRNA genes

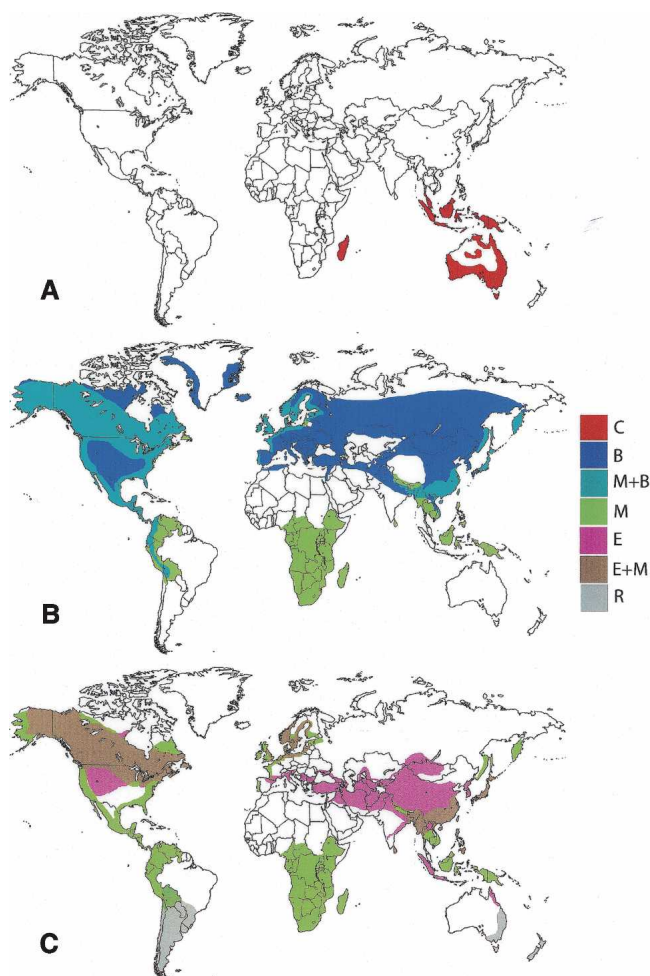


Figure 2. Present-day native distribution of actinorhizal plant hosts. (A) Distribution of plant hosts for CcI3, including *Casuarina* and *Allocasuarina* of the Casuarinaceae (C). (B) Distribution of plant hosts for ACN, including *Alnus* sp. in the Betulaceae (B) and Myricaceae (M) and their overlap (M+B). (C) Distribution of plant hosts for EAN including members of the Elaeagnaceae (E), Myricaceae (M), and the actinorhizal Tribe Colletieae of the Rhamnaceae in South America, Australia, and New Zealand (R). Elaeagnaceae and Myricaceae (E+M) overlap in some areas. Maps were drawn with information from Silvester (1977) and from the Missouri Botanical Garden Web site (www.mobot.org).

(Ochman et al. 1999), the clade containing EAN diverged an estimated 115 Mya from the clade containing CcI3 and ACN, which diverged from each other about 55 Mya (Fig. 1). Therefore, orthologous genes present in ACN and EAN but absent in CcI3 may be assumed to have been lost from CcI3 after it diverged from ACN. A similar approach can be used for identifying deleted genes from ACN. However, genes absent from EAN, but present in the other two strains, could either have been lost in EAN or acquired by horizontal gene transfer (HGT) in the progenitor of ACN and CcI3.

Defining the presence and absence of orthologs by using a reciprocal best BLAST hit criterion with an E-value threshold of 10^{-4} , we find that 1054 genes present in both ACN and EAN are not found in CcI3 (Table 3). Similarly, 466 genes are missing from ACN, and 555 are missing from EAN. The most logical explanation for the smaller number of genes missing in EAN and ACN versus CcI3 is an accelerated rate of gene loss in CcI3 over

the past 55 Myr since diverging from a common ancestor with ACN. Since EAN has had longer to lose genes, its apparent rate of loss is lower than that of either ACN or CcI3. The number deleted is underestimated in all cases since genes deleted from two of the three strains would appear as strain specific or horizontally transferred (below).

A BlastClust analysis was done to categorize deleted genes using 30% sequence identity over 52% of the length. This analysis indicated that some categories of genes were preferentially deleted over others. Genes annotated as hypothetical, conserved hypothetical, or of unknown function account for about one-third of the total (28%–34%) lost in each genome (Table 4). CcI3 is missing the most genes in all categories shown in Table 4 except for integrases and transposases; the latter have been especially reduced in ACN. In CcI3, genes involved in transport (including ABC transporters, solute-binding-dependent transport proteins), regulatory proteins (including genes in the categories TetR, IclR, LuxR, MarR, MerR), dehydrogenases, amidotransferase, oxygenases, and many hypothetical proteins, proteins of unknown function, and conserved hypothetical proteins have been lost. In short, genes encoding the capacity to scavenge and metabolize substrates from the environment have been reduced in CcI3.

Several genes lost by CcI3 are concerned with metabolic activities of potential importance to survival or symbiosis. These include genes encoding the DNA repair enzymes AP endonuclease, photolyase, DNA-formamidopyrimidine glycosylase, DNA alkylation repair, and RadC; two cellulases that might be involved in survival or infection; gas vesicle proteins whose loss could signal adaptation to dry environments where *Casuarina* sp. grow; general metabolism enzymes (NAD-dependent glutamate dehydrogenase, PEP carboxylase); and a large number of regulatory and solute transport proteins. Among the latter, there is only one iron siderophore gene cluster in CcI3 as compared with two in ACN and three in EAN. More directly related to symbiosis, CcI3 has lost one of the two copies of the *shc* (squalene hopene cyclase) genes involved in synthesizing bacteriohopane lipids that comprise the envelope of *Frankia* vesicles and provide protection for nitrogenase against oxygen. Unlike ACN and EAN, oxygen protection is conferred by secondary plant cell walls when CcI3 is in symbiosis (Berg and McDowell 1988), perhaps making bacteriohopane synthesis less of a priority.

In general, the classes of genes lost by CcI3 (DNA repair, metabolic enzymes, regulatory proteins) resemble those known to be lost by bacterial endosymbionts of animals (Mira et al.

Table 2. Summary of genome characteristics^a

	ACN (finished) ^b	CcI3 (finished) ^c	EAN (draft-1 scaffold) ^c
Size in bp	7,497,934	5,433,628	9,035,218
Predicted CDS	6786	4499	7976
Genes in COGs	4502 (67%)	2564 (57%)	4815 (60%)
tRNA	46	46	47
rRNA (5S-16S-23S)	2	2	3
%G+C	72.83	70.07	70.94
Accession	CT573213	CP000249	AAII00000000

^aNumbers are derived from the accession numbers indicated as of November 13, 2006.

^bNumbers are from <http://www.genoscope.cns.fr>.

^cNumbers are from accessions indicated at <http://www.ncbi.nlm.nih.gov> except for the genes in COGs for EAN, listed at <http://www.jgi.doe.gov>.

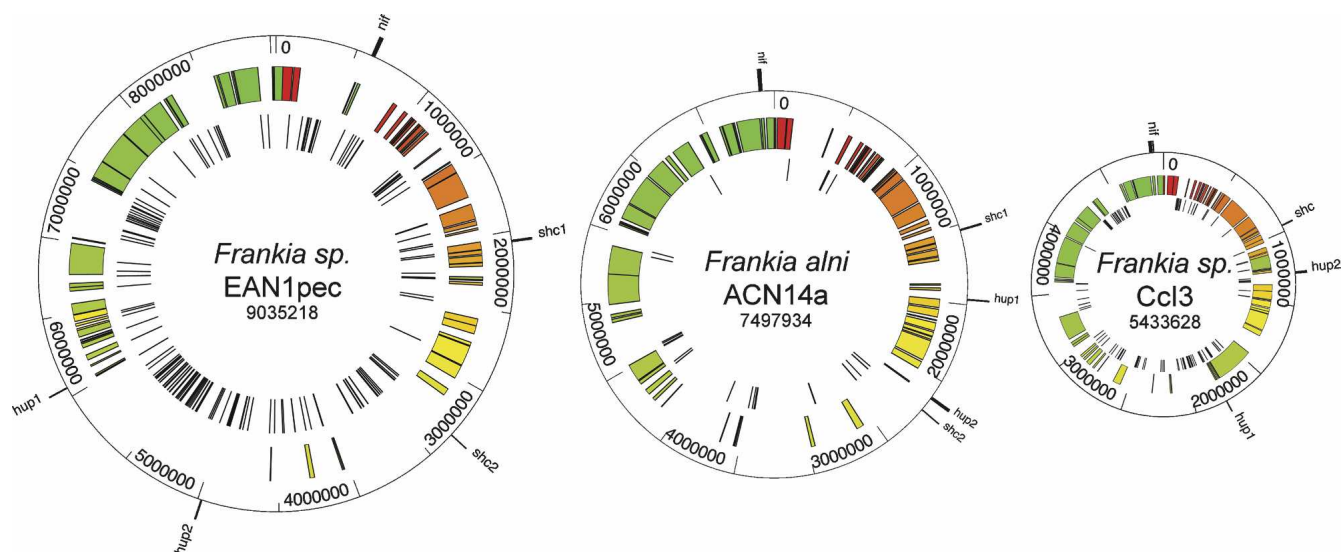


Figure 3. Genome maps of the three *Frankia* strains. Circles, from the outside in, show (1) gene regions related to symbiosis including *shc1*, *hup2*, *hup1*, and *nif*; (2) the coordinates in Mb beginning at 0 = oriC; (3) regions of synteny (syntons) calculated as a minimum of five contiguous genes present in all strains with an identity >30% over 80% of the length of the shortest gene (syntons are tagged with a spectrum-based [red-yellow-green] color code standardized on ACN to indicate regions where syntons have moved in the other strains); (4) IS elements and transposases. Circles were drawn using GenVision Software from DNASTar.

2001; Ochman and Moran 2001), and indicate that Ccl3 is evolving toward a greater dependence on its host. However, Ccl3 can still grow on minimal medium, so such strains have not yet been committed to an obligate symbiotic existence.

Gene duplication, acquisition, and ORFans

Gene duplication is a major means by which soil bacteria adapt to new niches, or to the availability of new substrates (Francino 2005; Konstantinidis and Tiedje 2005). Gene acquisition is known to be similarly involved in bacterial adaptation to new environments, particularly in the emergence of pathogens (Mira et al. 2001; Ochman and Moran 2001) and in the evolution of mutualistic bacteria in the legume symbiosis (Young and Haukka 1996; Chen et al. 2001; Moulin et al. 2004). *Frankia* symbionts have adapted both to living in diverse soils in most parts of the world and to living in root nodules from phylogenetically diverse angiosperms.

We defined duplicates as having the lowest BLAST E-value with a gene from the same genome when compared with genomes from other *Frankia* strains, *Acidothermus* and *Kineococcus*,

Table 3. Summary of genes involved in size differentiation of the three *Frankia* strains

Category	ACN	Ccl3	EAN
Deleted genes	466	1054	555
Duplicated genes	512 (7.5%)	444 (9.8%)	1355 (18.5%)
Strain-specific genes (SSGs)	709 (10.4%)	420 (9.3%)	934 (11.7%)
ORFans	854 (12.5%)	158 (3.5%)	355 (4.9%)
Transposases	33 (0.4%)	155 (3.4%)	195 (2.4%)
Integrases	13 (0.2%)	32 (0.7%)	74 (0.9%)

Deleted genes, duplicated genes, ORFans (no hits in any database), and strain-specific genes (SSG; hits in databases but not in the other two *Frankia* strains) were detected as described in the text. The number of genes annotated as transposases and inactivated derivatives plus integrases were annotated as described.

both close relatives to *Frankia* in the *Frankineae*, *Streptomyces* spp., and the NR (non-redundant) database. Using this approach, ~7.5% (512) of the ORFs in ACN, 9.8% (444) in Ccl3, and 18.5% (1355) in EAN could be considered duplicates of other genes in the same genomes (Table 3). Core metabolic genes are generally not duplicated, a differential amplification noted in other bacteria (Francino 2005; Konstantinidis and Tiedje 2005). Surprisingly, Ccl3, which has sustained strong reducing evolutionary pressures, nevertheless had a slightly higher percentage of duplicates than ACN, an observation that is accounted for by the proliferation of transposase genes in Ccl3 (Tables 3, 5). Gene duplication has thus enlarged the EAN genome to a greater ex-

Table 4. Categories of deleted genes assessed by BlastClust analysis

General category	Ccl3	ACN	EAN
Hypothetical, conserved hypothetical, unknown function	289	158	176
Transport-associated	113	33	28
Regulatory	95	21	35
Short-chain dehydrogenase/reductase	32	4	4
Acyl-CoA dehydrogenase-like	19	3	1
Alkanesulfonate monooxygenase	17	0	0
AMP-dependent synthetase and ligase	14	3	2
Protein kinase	12	2	6
Amidohydrolase	13	1	1
Enoyl-CoA hydratase/isomerase	10	1	2
L-carnitine dehydratase/bile acid inducible protein F	9	2	2
Alcohol dehydrogenase GroES-like	7	1	2
Cytochrome P450	8	5	3
Transposases	7	33	11
Integrases	3	7	6
Other	406	192	276

Deleted genes are defined as genes present in two of the three strains as assessed by BLAST hits below a cutoff of 10^{-4} but absent in the third. The general categories correspond to the major groups identified by a BlastClust analysis of missing genes.

Table 5. Top 20 families of duplicated genes in each *Frankia* strain assessed by BlastClust analysis

ACN		CcI3		EAN	
Annotated function	#	Annotated function	#	Annotated function	#
Putative serine/threonine protein kinase	22	Transposase, IS4	31	Short-chain dehydrogenase/reductase SDR Integrase, catalytic region	49
Putative oxidoreductase, short-chain dehydrogenase/reductase family	20	Transposase IS66	15		48
Hypothetical protein; putative HNH endonuclease domain	17	Transposase, IS4	14	ABC-type branched-chain amino acid transport, periplasmic component	32
Hypothetical protein; putative dehydrogenase	16	Transposase	12	Transposase, IS605 OrfB	27
Putative oligopeptide transport protein (ABC superfamily)	14	Regulatory protein, MerR:Recombinase	11	Acyl-CoA dehydrogenase-like	23
Putative SAM-dependent methyltransferases	13	Twin-arginine translocation pathway signal	7	Cytochrome P450	22
Putative alpha-methylacyl-CoA racemase	10	Transposase, IS4	7	Luciferase-like	21
Branched-chain amino acid ABC transport, binding protein	10	Transposase and inactivated derivatives-like	7	Integrase, catalytic region	20
Hypothetical protein; putative signal peptide	9	Transposase, IS111A/IS1328/IS1533:Transposase IS116/IS110/IS902	7	Amidohydrolase 2	20
Putative non-ribosomal peptide synthetase	8	Transposase (probable), IS891/IS1136/IS1341:Transposase, IS605 OrfB	7	Extracellular solute-binding protein, family 5	19
Branched-chain amino acid transport protein (ABC superfamily)	6	Putative IS630 family transposase	7	Membrane-bound lytic murein transglycosylase B-like	17
Conserved hypothetical protein; putative amidohydrolase domain	6	Hydantoinase/oxoprolinase	7	GGDEF domain	14
Putative GntR-family transcriptional regulator	6	ATP-binding region, ATPase-like	6	Putative transposase	13
Putative TetR-family transcriptional regulator	5	Putative O-methyltransferase	5	Transposase, IS4	13
Hypothetical protein; putative dibenzothiophene desulfurization	5	Putative plasmid replication initiator protein	5	ABC transporter related	12
Cytochrome P450	5	Transposase, IS4	4	Regulatory protein, LuxR	12
Hypothetical protein	5	Hypothetical protein	4	RNA-directed DNA polymerase	11
Putative monooxygenase	4	Amino acid adenylation	3	Phenylpropionate dioxygenase/related ring-hydroxylating dioxygenases	11
Hypothetical protein	4	Putative DNA-binding protein	3	Taurine catabolism dioxygenase TauD/TfdA	11
Putative aldehyde dehydrogenase	4	Hypothetical protein	3	Regulatory protein, TetR	11

The top 20 categories of duplicated genes were defined by a BlastClust analysis of a data set comprised of all duplicates. Duplicates were defined as having the best BLAST score of another gene within the same genome relative to genes within NR plus the other *Frankia* strains, *Acidothermus*, *Kineococcus*, and *Streptomyces* spp. with a maximum cutoff of 10^{-4} .

tent than the genomes of ACN or CcI3. Most of the duplicated genes seem to be located near the replication terminus in all strains. Localization of contingency genes to the terminus has been observed in the linear genomes of *Streptomyces* sp. and in other large genomes (Bentley et al. 2002; Ikeda et al. 2003).

To assess the types of genes duplicated, a BlastClust (NCBI) analysis was done to cluster proteins using a standard of 25% identity over at least 40% of the length of the amino acid sequence. A more stringent analysis using 30% identity over 52% of the sequence gave essentially the same results. In the top 20 duplicated gene families in CcI3, 116 out of 165 (70%) genes belong to several classes of transposases and genes associated with prophages (Table 5). In ACN, no transposases are found in the 151 genes in the top 20 families; instead, genes annotated as serine-threonine protein kinases, short-chain dehydrogenases/reductases, endonucleases, SAM-dependent methyltransferases, transport proteins, and a variety of dehydrogenases are duplicated. In EAN, 132 out of 406 (32.5%) genes are associated with

integrases, transposases, or reverse transcriptases in the top 20 families, with the remainder annotated as short-chain dehydrogenase/reductases, cytochrome P450s, transport proteins, regulatory proteins, and dioxygenases.

In sum, EAN has the most duplicated genes in all categories, including those whose products are associated with metabolic processes as well as mobile genetic elements. ACN has the fewest duplicates, and those are of genes involved in general metabolism. Finally, a large portion of all duplicates in CcI3 (33% overall) is of transposases. In all strains, the majority of duplicates appeared as two copies of a single gene.

Strain-specific genes (SSGs) include genes lost by two of the three *Frankia* strains plus genes that have no hits in databases (ORFans). Such genes could also have been horizontally transferred from other bacteria. Using a permissive threshold (E -value $\leq 10^{-4}$) between the genomes, and allowing self-genome hits to eliminate duplication, we found that ~23% (1563) of the genes in ACN, 12.8% (578) in CcI3, and 17.7% (1289) in EAN

have no clear homologs in the other two genomes. Of those, 854 (12.5%) in ACN, 158 (3.5%) in CcI3, and 355 (4.9%) in EAN were ORFans with no hits in NR, or the related *Acidothermus*, *Kineococcus*, or *Streptomyces* spp. genomes. The higher number and percentage of ORFans in ACN may reflect a lower evolutionary pressure to eliminate non-essential genes, a characteristic also reflected in its having the fewest deleted genes overall (Table 3).

Conclusions

We have shown that the unusual size divergence displayed by the *Frankia* genomes has arisen by the processes of deletion, duplication, and retention/acquisition operating in all strains but to different extents (Table 3). These processes have driven the genomes in different directions, reducing that of CcI3, expanding that of EAN, and keeping ACN relatively stable. The results of these broad comparisons lead us to propose a link between the biogeographic history of the actinorhizal plants and the genome evolution of the bacterial symbionts.

Evidence from ecological (Zimpfer et al. 1997), molecular ecological (Simonet et al. 1999), physiological (Sellstedt 1995), and now genomic studies indicates that *Casuarina* strains represented by CcI3 have evolved to become specialists with reduced genomes. Unlike *Alnus* and *Elaeagnus* strains, they have not been detected by trapping experiments in soils outside the native ranges of their host plants (Zimpfer et al. 1997; Simonet et al. 1999), and they infect a narrow spectrum of hosts (Fig. 2A; Table 1). Genome reduction is well documented in obligate pathogens and obligate symbionts in plants and animals (Mira et al. 2001; Ochman and Moran 2001; Moran 2003; Batut et al. 2004), and in some free-living cyanobacterial *Prochlorococcus* sp. (Dufresne et al. 2005). Genome reduction has not been described in bacterial facultative symbionts that also exist free-living in the soil; indeed, this is a most unexpected finding.

We suggest that a likely explanation for genome reduction in CcI3 is its geographic and symbiotic isolation in Australia and the Pacific islands, paralleling its host plants' isolation beginning ~100–65 Mya. *Casuarinaceae* species emerged as part of the flora of Gondwana as evidenced by fossils in New Zealand and South America that today are outside the native range (Campbell and Holden 1984). These plants, and their bacterial symbionts, coadapted to a hotter, drier climate as Australia split from Antarctica and moved north toward the equator. Present-day *Casuarina* strains live in locales where the soil biotic capacity is reduced and actinorhizal host diversity is limited.

In contrast, plants infected by *Elaeagnus* strains have a global distribution (Fig. 2C) with ancestral origins in both Gondwana (*Gymnostoma* in the *Casuarinaceae* in Western Oceania, actinorhizal *Colletieae* in the *Rhamnaceae*) and Laurasia (*Elaeagnaceae*, *Myricaceae*). Such plants occupy a wide range of soil types and climates. Genome expansion by gene duplication and divergence is a mechanism used by soil bacteria to exploit new niches and new substrates (Francino 2005; Konstantinidis and Tiedje 2005), and may be inferred to have occurred in the ancestors of EAN as they and their hosts coadapted to new and diverse soils. Indeed, the types of genes duplicated are largely involved in introducing substrates into central metabolic pathways.

The genome of ACN appears more stable than those of CcI3 and EAN, in the sense that it has few transposases and integrases; it also has lost the fewest genes by deletion, has the lowest proportion of duplicated genes, and retains the most strain-specific genes, including ORFans. Its stability may reflect its host range

focused on the ancient lineages in the *Betulaceae* and *Myricaceae* leading to high soil abundance and relatively strong genome homogenization. Its host plants have the longest fossil record of the N_2 -fixing clade (Magallon et al. 1999), and have inhabited similar and milder environments in northern latitudes since appearing in Laurasia during the late Cretaceous (Crane 1989).

Taken together, the gene contents of the three *Frankia* strains appear to reflect the biogeographic history of the host plants they infect, and as such may provide the first example of differential genome contraction and expansion occurring in closely related facultatively symbiotic soil bacteria that may be linked to the evolutionary history of their hosts on a global scale.

Methods

Strains

CcI3 was isolated from *Casuarina cunninghamiana* plants growing in a greenhouse at Harvard Forest in Petersham, MA (Zhang et al. 1984) on soils coming from its original provenance. ACN was isolated initially from *Alnus viridis* subsp. *crispa* plants in Tadousac, Quebec (Normand and Lalonde 1982; Benson et al. 2004). Strain EAN was isolated from field nodules of *E. angustifolia* growing in Ohio (Lalonde et al. 1981).

Genome sequencing, assembly, and finishing—CcI3 and EAN1pec

We sequenced the three genomes of *Frankia* strains ACN14a, CcI3, and EAN1pec using a shotgun approach. The genomes of *Frankia* strains CcI3 and EAN1pec were sequenced at the Joint Genome Institute (JGI) using a combination of 3-kb, 8-kb, and 40-kb (fosmid) DNA libraries for each strain. Draft assemblies were based on 82,561 total reads for CcI3 and 125,615 total reads for EAN1pec. The different libraries provided $4.6 \times$ (3-kb), $4.1 \times$ (8-kb), and $0.5 \times$ (fosmids) coverage of CcI3 and $4.0 \times$ (3-kb), $3.4 \times$ (8-kb), and $0.6 \times$ (fosmids) coverage of EAN1pec. End sequencing and fingerprinting of fosmid clones aided in assembly verification, determination of gap sizes, and ordering and orientation of scaffolds beyond assembly gaps.

Sequencing gaps were closed mainly by primer walking on plasmid and fosmid subclone templates. In cases where no acceptable template was available, PCR products were made and sequenced using customized primers. Gaps resulting from hard-to-sequence DNA structures had to be covered using special chemistries and protocols developed in-house. Mis-assemblies were identified and corrected by means of clone pairing; these primarily occurred due to long repeats (rRNAs, IS elements). Over-collapsing of repeat copies often resulted in pseudo-gaps in the assembly, which could not be closed by routine primer walking. Each one of those had to be filled in using one of the following two methods. Small pseudo-gaps were closed using the editing features of CONSED (Gordon et al. 1998), by locating and placing appropriate reads individually into their proper repeat copy. Long pseudo-gaps and long misassembled repeats (>2 kb) had to be isolated and separately assembled. Only consistent, partially unique clone-mates would be allowed in those subassemblies. After verifying the subassembly's integrity and primer-walking over the poorly covered regions, the isolated contigs were reintroduced into the main assembly as "fake reads," that is, single continuous long sequences reflecting the correctly assembled repeat copy.

All other general aspects of library construction, sequencing, and automated annotation were carried out as previously de-

scribed for bacterial genomes sequenced at the JGI (Chain et al. 2003). In addition, predicted coding sequences are subject to manual analysis using the Integrated Microbial Genomes (IMG) annotation pipeline. Detailed information about genome annotation and other genome properties can be obtained at <http://img.jgi.doe.gov> (Markowitz et al. 2006).

Genome sequencing, assembly, and annotation—ACN14a

For ACN14a, four libraries were made: Two plasmid libraries of 3 kb and 10 kb, obtained by mechanical shearing, were constructed at Genoscope (Evry, France) into pCDNA2.1 (Invitrogen) and into the pCNS home vector (pSU18 modified, Bartolome et al. 1991), respectively. Two BAC libraries of an average insert size of 104 kb were constructed at Clemson University Genomics Institute (CUGI) by enzymatic digestion (EcoRI and HindIII) into pCUGIBAC1 (Luo and Wing 2003). Plasmid and BAC DNAs were purified and end-sequenced using dye-terminator chemistry on ABI3730xl DNA Analyzer sequencers. We generated 150,890 sequences from both ends of genomic clones from the four libraries.

The Phred/Phrap/Consed software package (www.phrap.com) was used for sequence assembly and quality assessment (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). A total of 8956 additional reactions were necessary to close gaps, generally with the transposing method (Entranceposon, Finnzymes, Espoo, Finland) and to raise the quality of the finished sequence. The validity of the final sequence was assessed by comparing the restriction enzyme pattern deduced from the sequence to the experimentally observed restriction pattern obtained by digestion of genomic DNA. After a first round of annotation, regions of lower quality as well as regions with putative frame-shifts were resequenced from PCR amplification of the dubious regions.

Using the AMIGene software (Bocs et al. 2003), a total of 5279 CDSs were predicted and submitted to automatic functional annotation (Vallenet et al. 2006). Each predicted gene was assigned a unique identifier prefixed with "FRAAL." Sequence data for comparative analyses were obtained from the NCBI databank (RefSeq section). Putative orthologs and synteny groups (conservation of the chromosomal co-localization between pairs of orthologous genes from different genomes) were computed between ACN and the 265 other complete genomes using the procedure described in Vallenet et al. (2006). Manual validation of the automatic annotation was performed using the MaGe (Magnifying Genomes) interface, which allows graphic visualization of the ACN annotations enhanced by a synchronized representation of synteny groups in other genomes chosen for comparisons. The ACN nucleotide sequence and annotation data have been deposited at EMBL databank under accession number CT573213. In addition, all the data (i.e., syntactic and functional annotations, and results of comparative analysis) were stored in a relational database, called FrankiaScope (Vallenet et al. 2006). This database is publicly available via the MaGe interface at <http://www.genoscope.cns.fr/agc/mage/frankia/Login/log.php>.

Methods used for determining deleted, duplicated, strain-specific, and ORFan genes

Genes deleted from one strain were identified by using reciprocal BLAST hits from each pair of genomes. That is, each pair of orthologs identified each other as the lowest BLAST score. Genes were scored as deleted if they did not have a reciprocal hit in another *Frankia* genome. The data set included three *Frankia* strains.

Gene duplications were assessed as having the best BLAST

hits within the same genome (duplicates) using an E-value cutoff of 10^{-4} and a data set consisting of NR (minus *Frankia* sequences) + *Kineococcus radiodurans* + *Streptomyces coelicolor* + *S. avermitilis* + ACN + Cc13 + EAN. To cluster duplicates, the program BlastClust (NCBI) was used with settings reported in the text.

Strain-specific genes (SSGs) include genes found in one but not another *Frankia* strain at an E-value cutoff of 10^{-4} , plus genes that have no hits in databases. The latter are referred to as ORFans.

Acknowledgments

This work was supported by the National Science Foundation Microbial Genome sequencing program to D.R.B., L.S.T., and M.P.F. The work on Cc13 and EAN was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological, and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory. The work on ACN14a was performed at Genoscope, Evry, France and was supported by CNRS/ACI Microbiologie and MRT/ACI IMPBio2004. Funds to Anita Sellstedt are from FORMAS and NFR, Sweden. Funds for the construction of the ACN BAC library were provided in part by the Genome Science and Technology Program, The University of Tennessee. Correspondence and requests for materials should be addressed to D.R.B. (david.benson@uconn.edu) for Cc13, to P.N. (normand@biomserv.univ-lyon1.fr) for ACN, and to L.S.T. (Ist@hypatia.unh.edu) for EAN.

References

- The Angiosperm Phylogeny Group (APG). 1998. An ordinal classification for the families of flowering plants. *Ann. Mo. Bot. Gard.* **85**: 531–553.
- Bartolome, B., Jubete, Y., Martinez, E., and de la Cruz, F. 1991. Construction and properties of a family of pACYC184-derived cloning vectors compatible with pBR322 and its derivatives. *Gene* **102**: 75–78.
- Batut, J., Andersson, S.G., and O'Callaghan, D. 2004. The evolution of chronic infection strategies in the α -proteobacteria. *Nat. Rev. Microbiol.* **2**: 933–945.
- Benson, D.R., Vanden Heuvel, B.D., and Potter, D. 2004. Actinorhizal symbioses: Diversity and biogeography. In *Plant microbiology* (ed. M. Gillings). BIOS Scientific Publishers Ltd., Oxford.
- Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., et al. 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**: 141–147.
- Berg, R.H. and McDowell, L. 1988. Cytochemistry of the wall of infected *Casuarina* actinorhizae. *Can. J. Bot.* **66**: 2038–2047.
- Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G., and Medigue, C. 2003. AMIGene: Annotation of Microbial Genes. *Nucleic Acids Res.* **31**: 3723–3726.
- Campbell, J.D. and Holden, A.M. 1984. Miocene casuarinacean fossils from Southland and Central Otago, New Zealand. *N.Z. J. Bot.* **22**: 159–167.
- Chain, P., Kurtz, S., Ohlebusch, E., and Slezak, T. 2003. An applications-focused review of comparative genomics tools: Capabilities, limitations and future challenges. *Brief. Bioinform.* **4**: 105–123.
- Chen, L.-M., Cui, Y., Qin, M., Wang, Y., Bai, X., and Ma, Q. 1991. Identification of a *nodD*-like gene in *Frankia* by direct complementation of a *Rhizobium nodD*-mutant. *Mol. Gen. Genet.* **233**: 311–314.
- Chen, W.M., Laevens, S., Lee, T.M., Coenye, T., De Vos, P., Mergeay, M., and Vandamme, P. 2001. *Ralstonia taiwanensis* sp. nov., isolated from root nodules of *Mimosa* species and sputum of a cystic fibrosis patient. *Int. J. Syst. Evol. Microbiol.* **51**: 1729–1735.
- Clawson, M.L., Bourret, A., and Benson, D.R. 2004. Assessing the phylogeny of *Frankia*-actinorhizal plant nitrogen-fixing root nodule symbioses with *Frankia* 16S rRNA and glutamine synthetase gene sequences. *Mol. Phylogenet. Evol.* **31**: 131–138.

- Crane, P.R. 1989. Early fossil history and evolution of the Betulaceae. In *Evolution, systematics, and fossil history of the Hamamelidae* (eds. P.R. Crane and S. Blackmore), pp. 87–116. Clarendon Press, Oxford.
- Dufresne, A., Garczarek, L., and Partensky, F. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* **6**: R14.
- Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**: 186–194.
- Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**: 175–185.
- Francino, M.P. 2005. An adaptive radiation model for the origin of new gene functions. *Nat. Genet.* **37**: 573–577.
- Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. 2005. Mobile genetic elements: The agents of open source evolution. *Nat. Rev. Microbiol.* **3**: 722–732.
- Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F.L., et al. 2005. Opinion: Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3**: 733–739.
- Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8**: 195–202.
- Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M., and Omura, S. 2003. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat. Biotechnol.* **21**: 526–531.
- Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**: 111–120.
- Konstantinidis, K.T. and Tiedje, J.M. 2005. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci.* **102**: 2567–2572.
- Lalonde, M., Calvert, H.E., and Pine, S. 1981. Isolation and use of *Frankia* strains in actinorhizae formation. In *Current perspectives in nitrogen fixation* (eds. A.H. Gibson and W.E. Newton), pp. 296–299. Australian Academy of Sciences, Canberra.
- Luo, M. and Wing, R.A. 2003. An improved method for plant BAC library construction. *Methods Mol. Biol.* **236**: 3–20.
- Magallon, S., Crane, P.R., and Herendeen, P.S. 1999. Phylogenetic pattern, diversity, and diversification of eudicots. *Ann. Mo. Bot. Gard.* **86**: 297–372.
- Markowitz, V., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., et al. 2006. The Integrated Microbial Genomes (IMG) system. *Nucleic Acids Res.* **34**: D344–D348.
- Mira, A., Ochman, H., and Moran, N.A. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**: 589–596.
- Moran, N.A. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr. Opin. Microbiol.* **6**: 512–518.
- Moulin, L., Munive, A., Dreyfus, B., and Boivin-Masson, C. 2001. Nodulation of legumes by members of the β -subclass of Proteobacteria. *Nature* **411**: 948–950.
- Moulin, L., Bena, G., Boivin-Masson, C., and Stepkowski, T. 2004. Phylogenetic analyses of symbiotic nodulation genes support vertical and lateral gene co-transfer within the *Bradyrhizobium* genus. *Mol. Phylogenet. Evol.* **30**: 720–732.
- Normand, P. and Fernandez, M.P. 2007. Evolution and diversity of *Frankia*. In *Actinorhizal symbioses* (ed. K. Pawlowski). Springer Verlag, London (in press).
- Normand, P. and Lalonde, M. 1982. Evaluation of *Frankia* strains isolated from provenances of two *Alnus* species. *Can. J. Microbiol.* **28**: 1133–1142.
- Normand, P., Orso, S., Cournoyer, B., Jeannin, P., Chapelon, C., Dawson, J., Evtushenko, L., and Misra, A.K. 1996. Molecular phylogeny of the genus *Frankia* and related genera and emendation of family *Frankiaceae*. *Int. J. Syst. Bacteriol.* **46**: 1–9.
- Ochman, H. and Moran, N.A. 2001. Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science* **292**: 1096–1099.
- Ochman, H., Elwyn, S., and Moran, N.A. 1999. Calibrating bacterial evolution. *Proc. Natl. Acad. Sci.* **96**: 12638–12643.
- Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406–425.
- Sellstedt, A. 1995. Specificity and effectivity in nodulation by *Frankia* on southern hemisphere actinorhiza. *FEMS Microbiol. Lett.* **125**: 231–236.
- Silvester, W.B. 1976. Ecological and economic significance of the non-legume symbiosis. In *Proceedings of the First International Symposium on Nitrogen Fixation* (eds. W.E. Newton and C.J. Nyman), pp. 489–506. Washington University Press, Pullman, Washington, D.C.
- Silvester, W.B. 1977. Dinitrogen fixation by plant associations excluding legumes. In *A Treatise on dinitrogen fixation* (eds. R.W.F. Hardy and A.H. Gibson), pp. 141–190. John Wiley and Sons, New York.
- Simonet, P., Navarro, E., Rouvier, C., Reddell, P., Zimpfer, J., Dommergues, Y., Bardin, R., Combarro, P., Hamelin, J., Domenach, A.-M., et al. 1999. Co-evolution between *Frankia* populations and host plants in the family Casuarinaceae and consequent patterns of global dispersal. *Environ. Microbiol.* **1**: 525–533.
- Soltis, D.E., Soltis, P.S., Morgan, D.R., Swensen, S.M., Mullin, B.C., Dowd, J.M., and Martin, P.G. 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proc. Natl. Acad. Sci.* **92**: 2647–2651.
- Swensen, S.M. 1996. The evolution of actinorhizal symbioses: Evidence for multiple origins of the symbiotic association. *Am. J. Bot.* **83**: 1503–1512.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**: 4876–4882.
- Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., and Médigue, C. 2006. MaGe—A microbial genome annotation system supported by synteny results. *Nucleic Acids Res.* **34**: 53–65.
- Vanden Heuvel, B.D., Benson, D.R., Bortiri, E., and Potter, D. 2004. Low genetic diversity among *Frankia* spp. strains nodulating sympatric populations of actinorhizal species of Rosaceae, *Ceanothus* (Rhamnaceae) and *Datisca glomerata* (Datisceae) west of the Sierra Nevada (California). *Can. J. Microbiol.* **50**: 989–1000.
- Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., et al. 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* **37**: 463–464.
- Young, J.P. and Haukka, K.E. 1996. Diversity and phylogeny of rhizobia. *New Phytol.* **133**: 87–94.
- Zhang, Z., Lopez, M.F., and Torrey, J.G. 1984. A comparison of cultural characteristics and infectivity of *Frankia* isolates from root nodules of *Casuarina* species. *Plant Soil* **78**: 79–90.
- Zimpfer, J.F., Smyth, C.A., and Dawson, J.O. 1997. The capacity of Jamaican mine spoils, agricultural and forest soils to nodulate *Myrica cerifera*, *Leucaena leucocephala* and *Casuarina cunninghamiana*. *Physiol. Plant.* **99**: 664–672.

Received August 1, 2006; accepted in revised form October 18, 2006.